# Salivary Proteome Wiki: A Collaborative Environment to Share and Annotate Proteomics Data

# I204

William W. Lau[1], Lillian Shum, PhD[2], Calvin A. Johnson, PhD[1]

[1]Division of Computational Bioscience, Center for Information Technology, [2] Division of Extramural Research, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD, United States

## Introduction

In March 2008, a consortium of researchers across the United States completed a catalog of 1,116 proteins present in human saliva [1]. Tandem mass spectrometry in combination with liquid chromatography and electrospraYionization were used to generate mass spectra of peptide fragments. Proteins were then identified by matching the list of observed fragment masses with predicted masses in either SEQUEST or Mascot.

The use of saliva as a diagnostic fluid offers many advantages over traditional blood-based tests. To facilitate the identification of biomarkers with diagnostic values in saliva, the salivary proteome catalog has been made available online in an interactive Web portal known as the Salivary Proteome Wiki (SPW). In addition to browsing the catalog, users can add their own research data; share results; curate the data; and discover new knowledge. The goal of this resource is to provide the research community a convenient location with all the necessary tools in place to further enrich and refine the catalog.

## Methods

Many public resources, such as PRIDE [2] and Human Proteinpedia [3], have been developed as central repositories for proteomics data. Users can provide annotations on data that they submitted to enable primarily data sharing and searching. On the other hand, the Salivary Proteome Wiki uses a community-oriented approach, where users can perform an array of functions depending on their roles.

### Data Sources

Currently, more than 170 experiments, 2300 protein identifications, and 175,000 peptide hits are stored in the system as wiki pages (Fig. 1). Each page consists of annotations made up of free text or controlled vocabulary terms, from Gene Ontology for instance, to describe various properties of the entity. Users with the appropriate level of permission can add new annotations or edit existing ones in a form-based interface (Fig. 2). They can select annotation terms by browsing the vocabulary through an ontology lookup tool. On protein pages, the system also provides users a facility to check the local data against the International Protein Index (IPI) database to determine whether any updates are needed. Any proposed changes will have to go through a review process, in which other users can give feedback to the proposal and approval from a curator is needed before the changes become official. Alternatively, a discussion area is available on each page for users to ask questions, give suggestions, and exchange ideas in a less formal, uncurated setting.

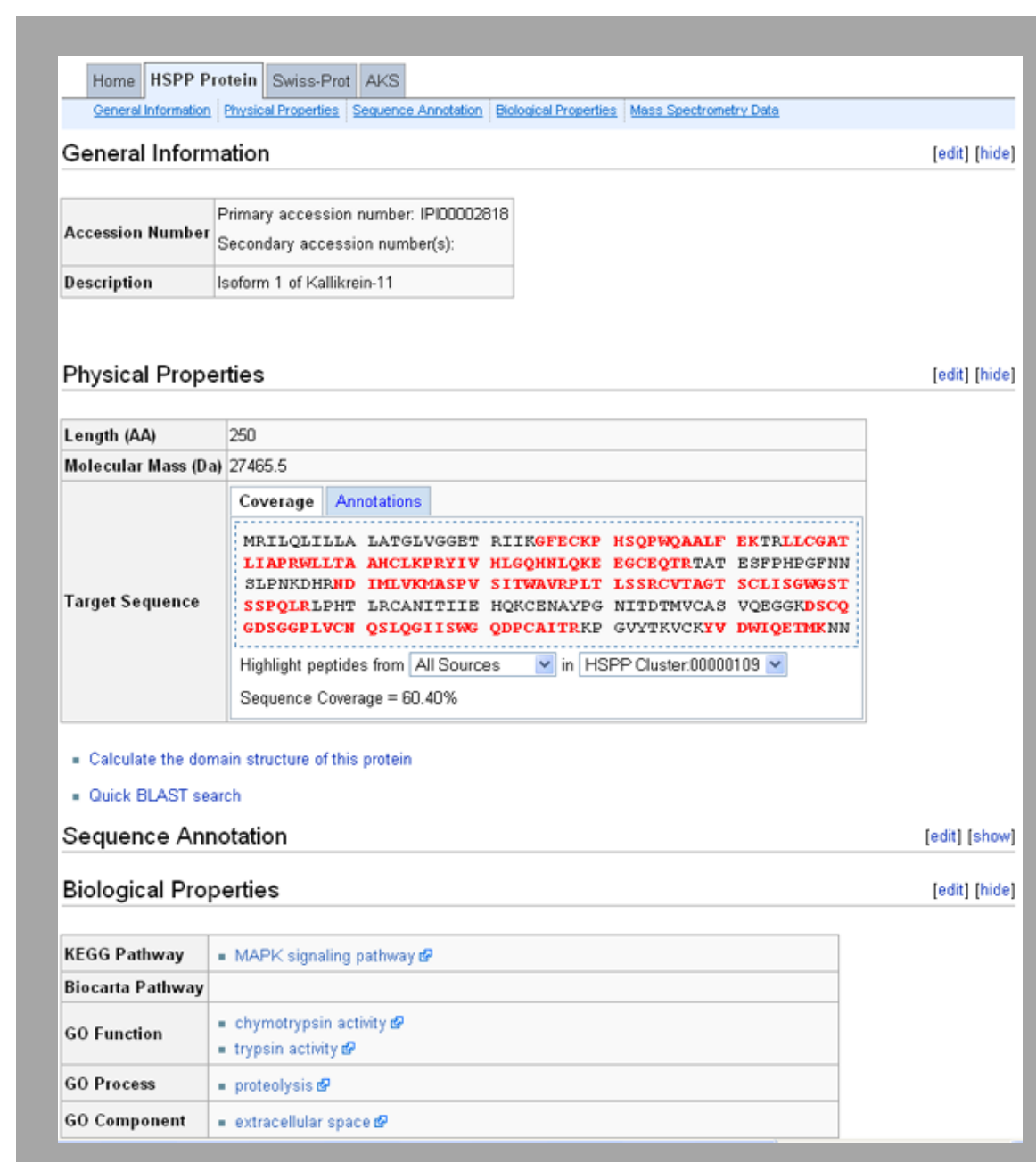


**FIGURE 1.** The protein page for Isoform 1 of Kallikrein-11, showing various properties of the protein and its amino acid sequence with peptides identified from experiments highlighted.

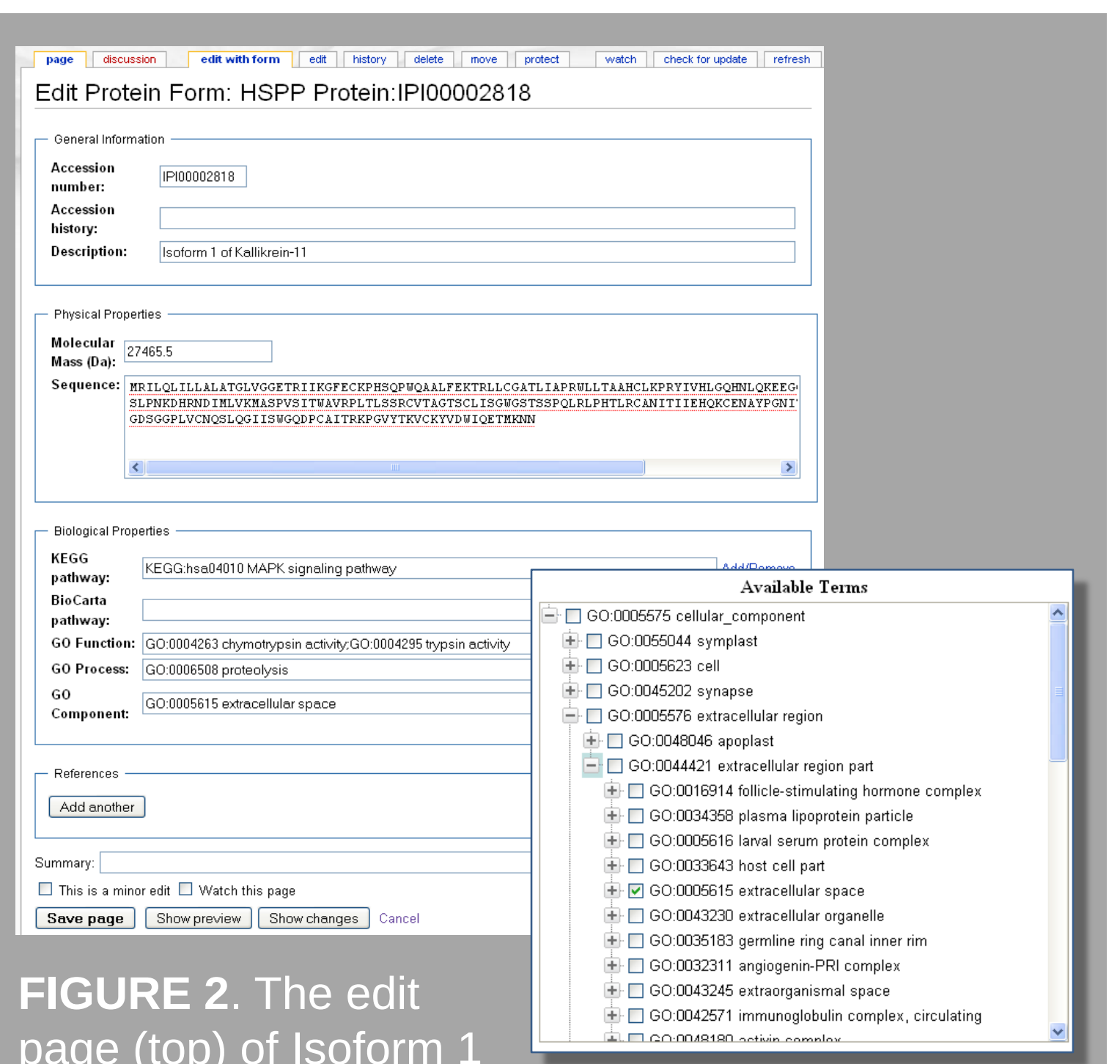


**FIGURE 2.** The edit page (top) of Isoform 1 of Kallikrein-11, where users can modify annotations using free-text or by selecting controlled vocabulary terms from a pop-up menu (inset) listing all available choices hierarchically.

In addition to the data generated from the Human Salivary Proteome Project, a number of popular knowledge bases, including InterPro, PubMed, and Swiss-Prot, have been incorporated into the system as well. Data from different sources are linked automatically through accession mapping to let users quickly gather all the relevant information.

### Knowledge Discovery Infrastructure

Wikis are largely text-based systems. To make the system aware of the meaning of the information stored in the database, a large amount of data are marked up semantically to enable ad-hoc querying and dynamic linking of related pages (Fig. 3). Each semantic annotation is a triplet, composed of a subject, predicate, and the value. One such annotation in the Kallikrein-11 protein page is *<HSPP Protein:IPI00002818, participates in, GO:0006508 proteolysis>*. One potential use of these triplets is to utilize Semantic Web technologies to perform machine-based knowledge discovery.

Another discovery feature is a concept extraction tool that has been implemented to identify biomedical terms in free text (Fig 4.). Each term is mapped to a specific concept in the ontology browser, where users can learn more about the term, including its definitions, synonyms, and other related concepts.

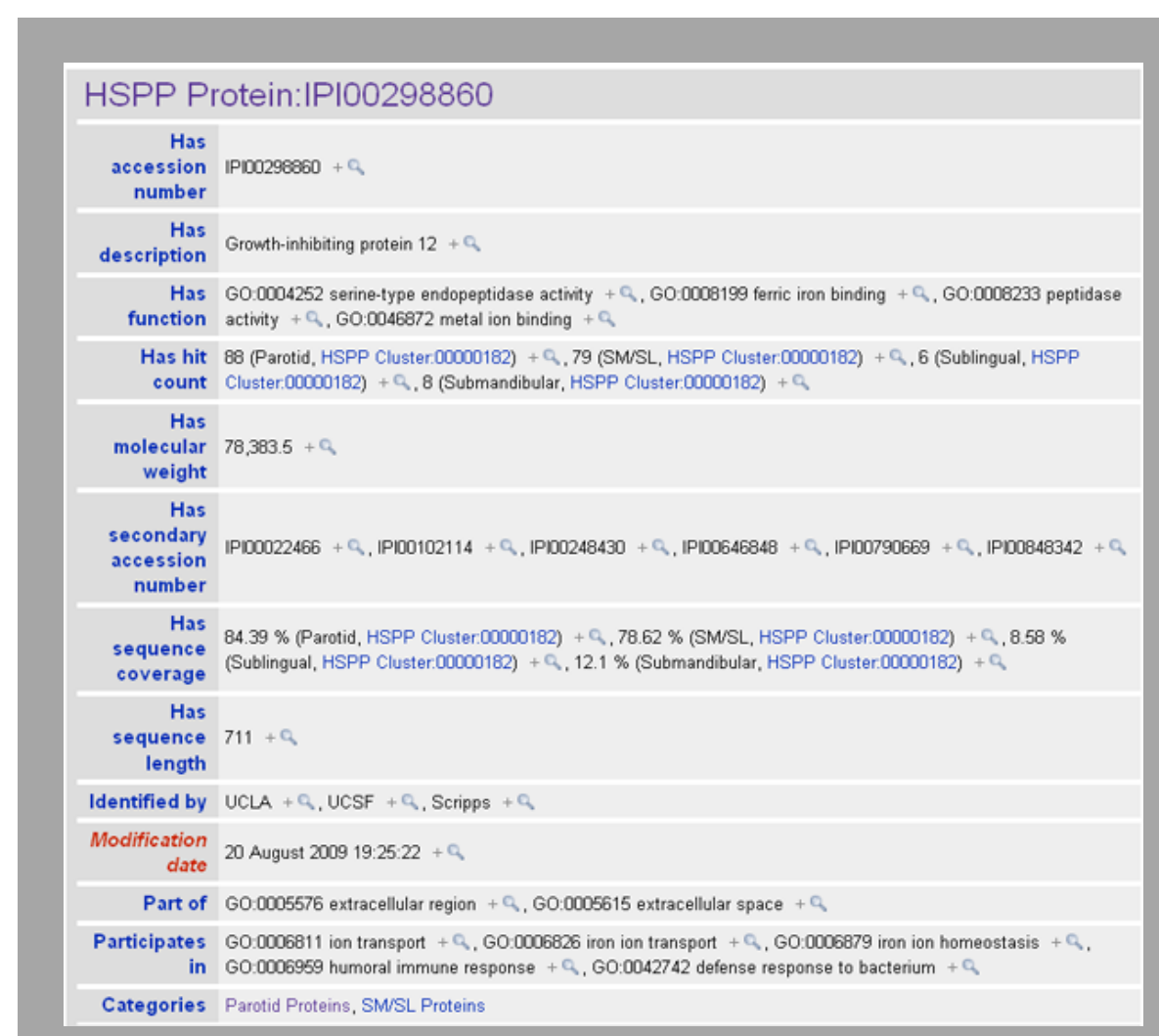


**FIGURE 3.** A sample list of protein properties that are semantically annotated by the system using Semantic Web data models to facilitate advanced searching and data exchange with other systems.

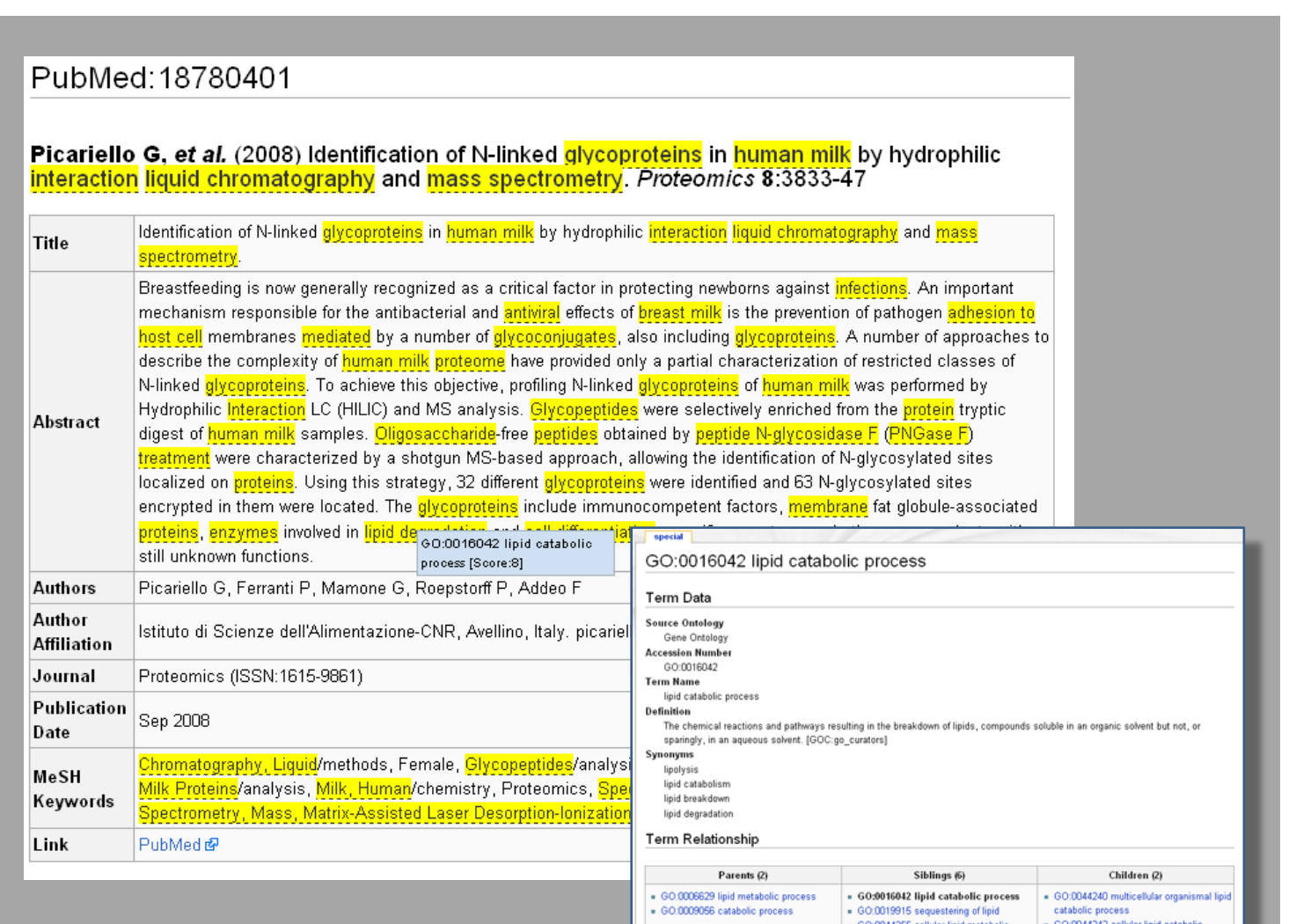


**FIGURE 4.** A PubMed page (top) with the concept extraction tool turned on to highlight biomedical concepts. Clicking on a term will direct users to the ontology term page (inset) with additional information about the concept.

A commercial knowledge discovery engine, known as AKS$^2$ (Alma Bioinformatics, Madrid, Spain) [4], has also been integrated into the wiki to support exploration of biomedical concepts (Figs. 5-7). Users can launch AKS$^2$ through direct links from the protein pages or from a search box. Within the tool, they can:

1.Visualize publication trends and relationship networks;
2.highlight concepts in literature;
3.extract relations between biomedical concepts;
4.identify experts in an area of research; and
5.group related concepts into clusters based on the strength of their relationship.

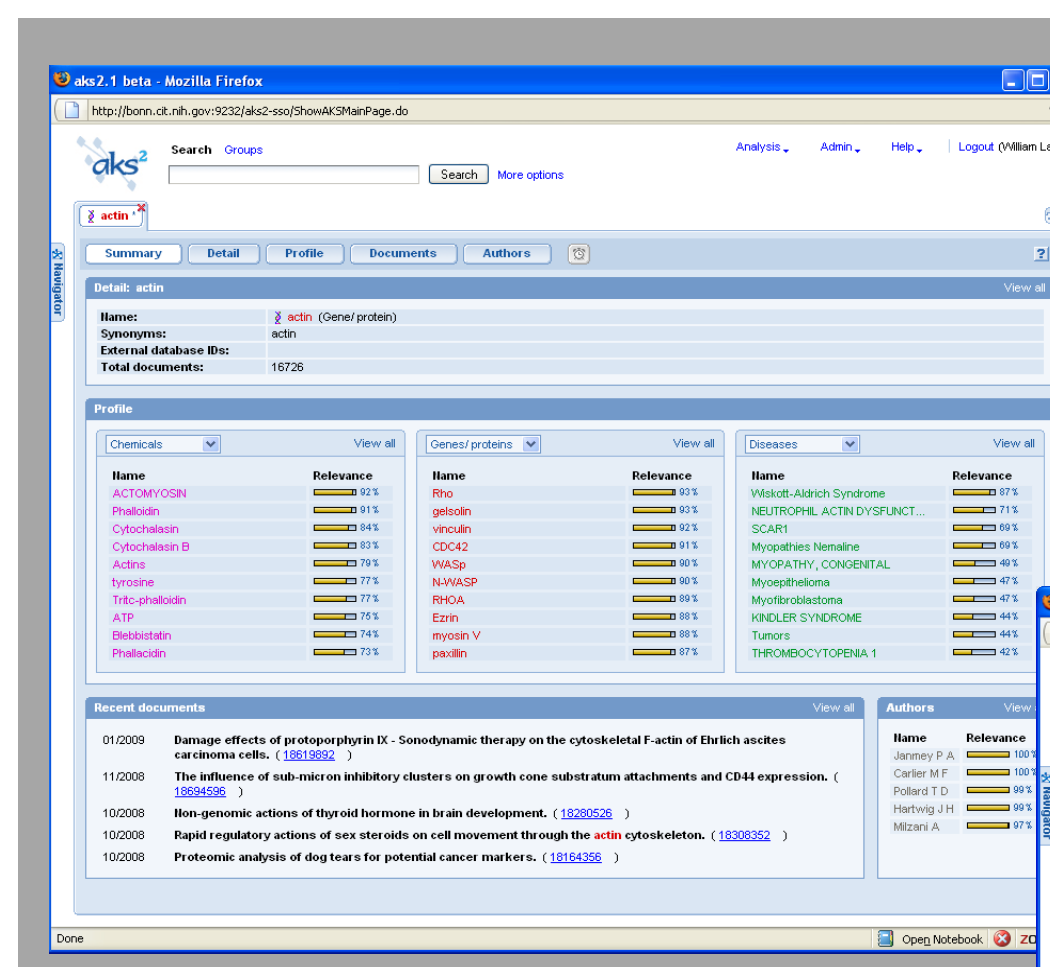


**FIGURE 5.** An AKS$^2$ interface showing a summary of the information available for the protein actin, including literature, experts, and relations with other concepts.

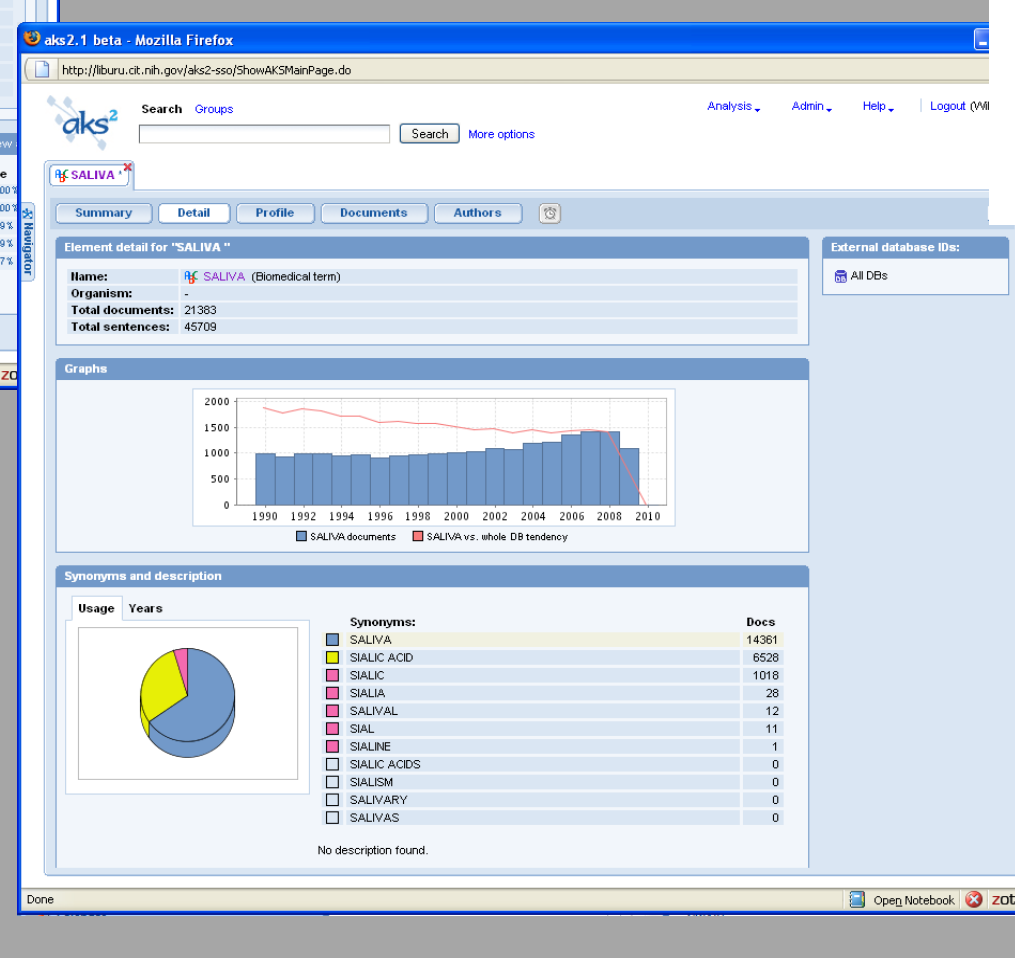


**FIGURE 6.** Publication trend of saliva-related literature and usage information of the different synonyms of the concept.

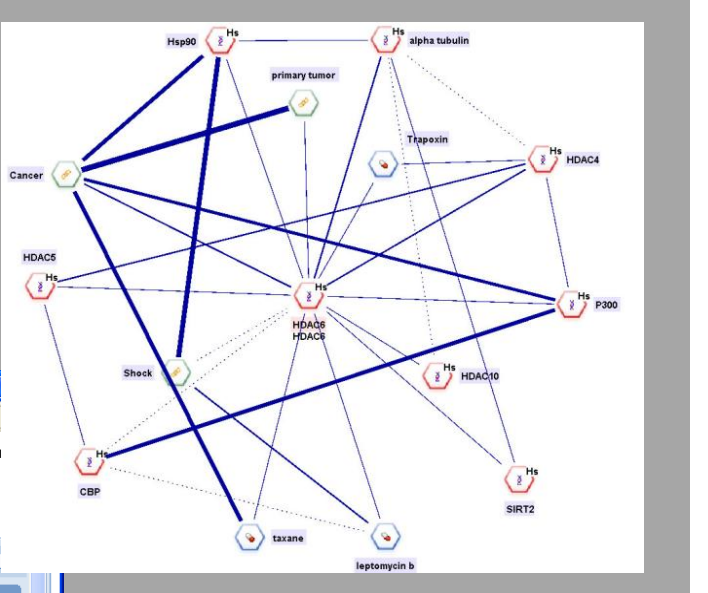
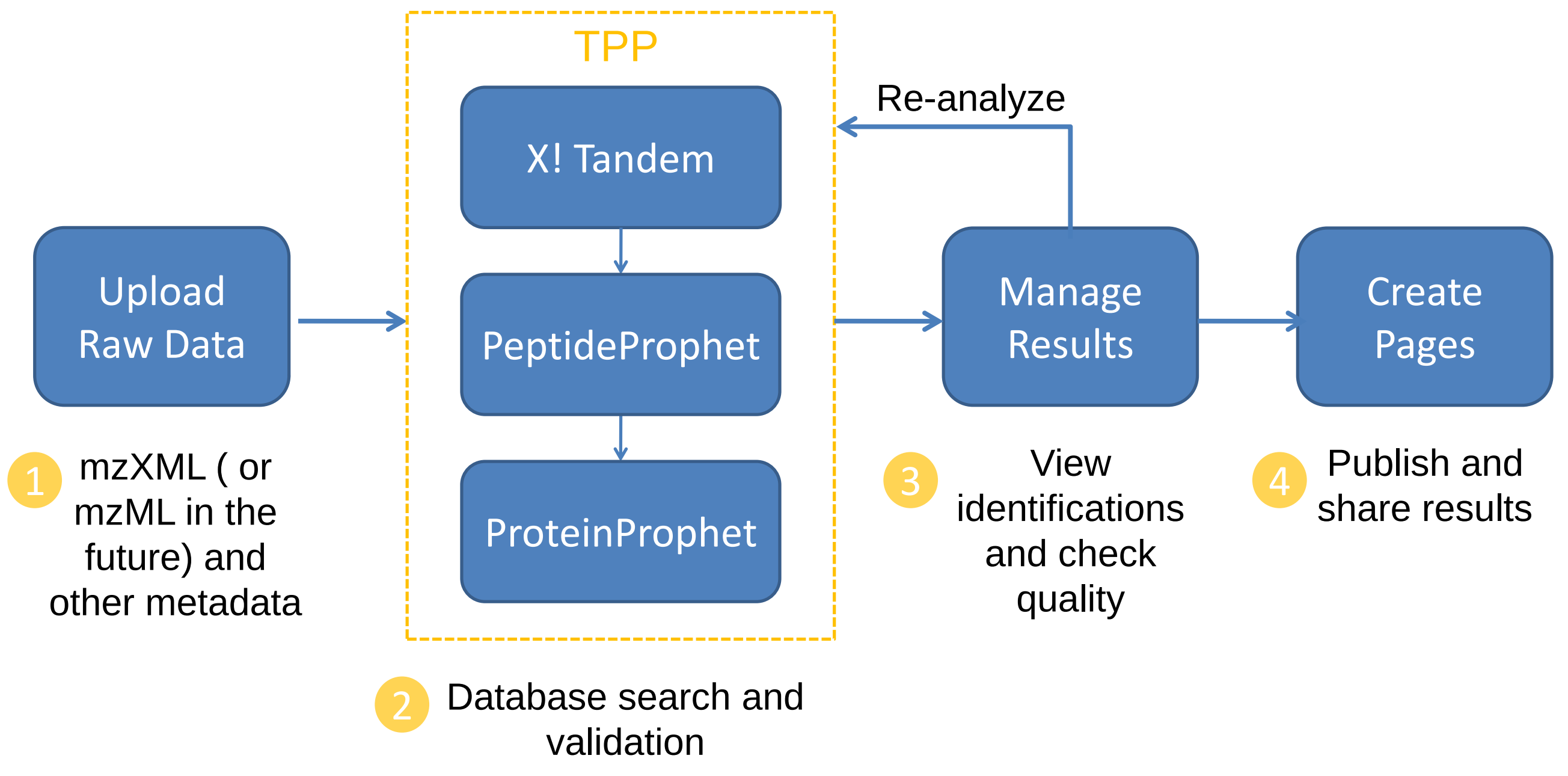


**FIGURE 7.** A network graph illustrating the strength of connection among related concepts.

### Proteomics Data Analysis Pipeline

The proteomics data analysis pipeline in the SPW is an implementation of the widely-used Trans-Proteomics Pipeline (TPP) software developed at the Institute for Systems Biology [5]. The pipeline allows users to upload the results of a MS/MS experiment in mzXML format, perform the typical steps of identification and validation, and subsequently publish the findings as wiki pages (Fig. 8). Once published, all the associated data, including the spectra (Fig. 9) and the identifications (Fig. 10), will be available to other users of the system, who can then reproduce the experiment and verify the results.



TPP

Upload Raw Data

X! Tandem

PeptideProphet

ProteinProphet

Manage Results

Re-analyze

Create Pages

1 mzXML ( or mzML in the future) and other metadata

2 Database search and validation

3 View identifications and check quality

4 Publish and share results

**FIGURE 8.** A diagram showing the steps from uploading raw data to publishing results in the proteomics data analysis pipeline.

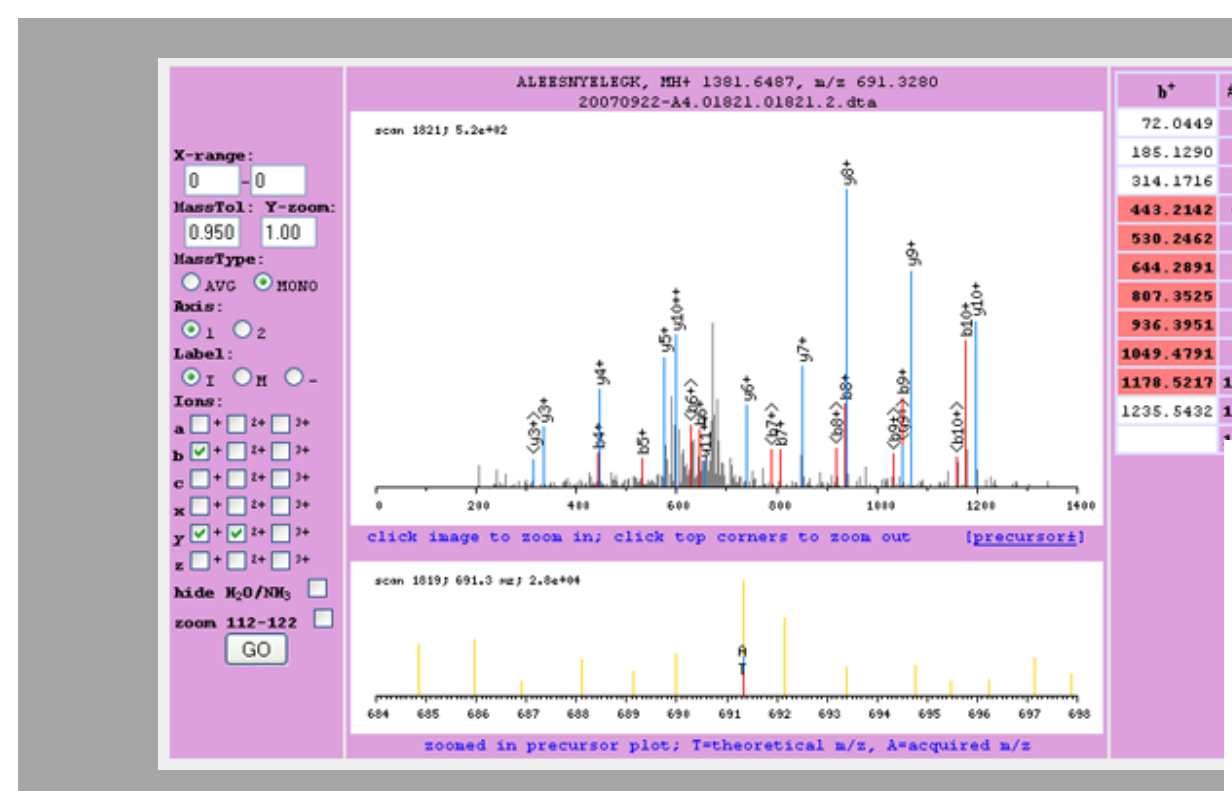


**FIGURE 9.** The COMET Spectrum View of a peptide identified in an experiment is a feature of the Trans-Proteomics Pipeline (TPP).

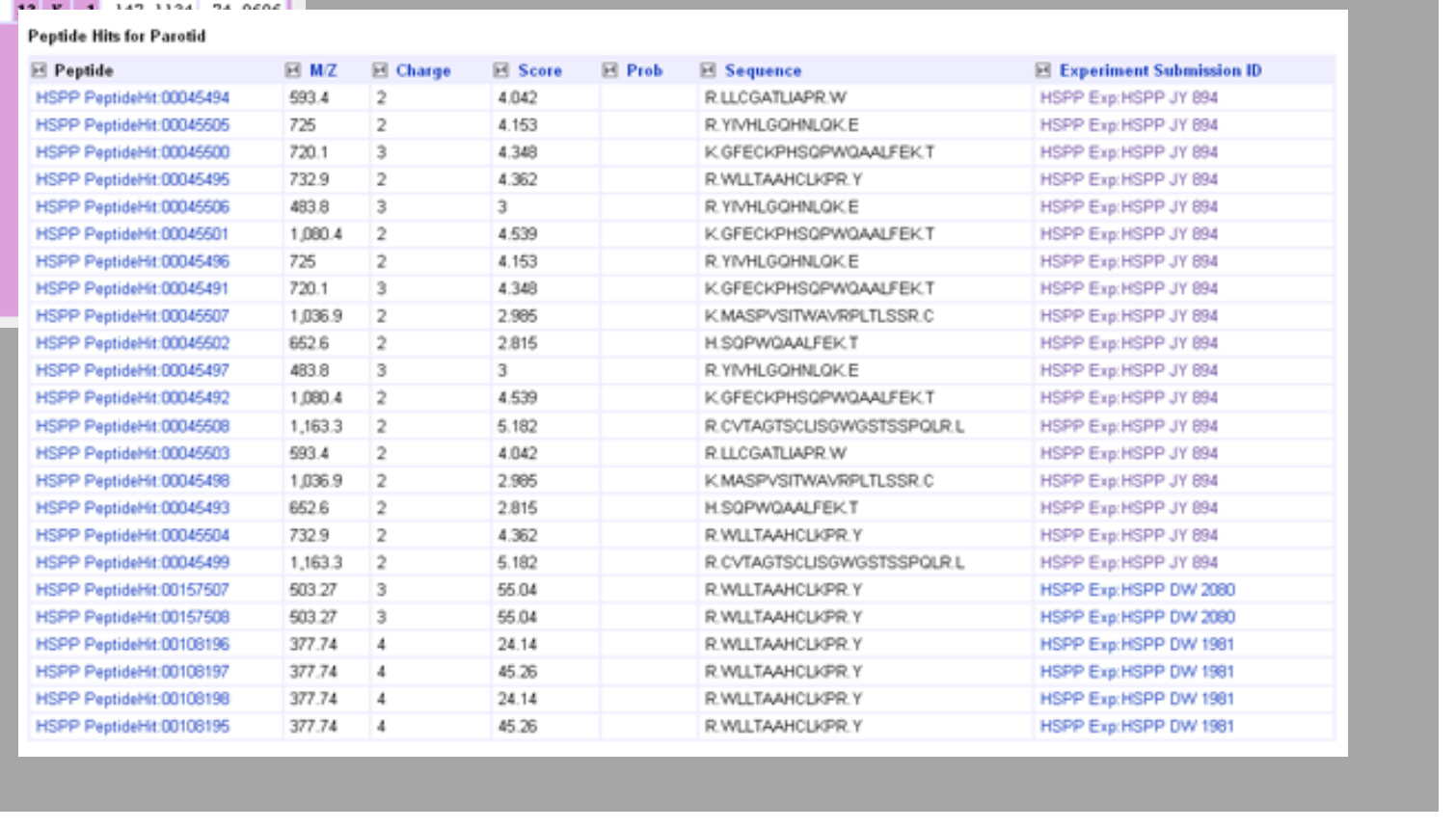


**FIGURE 10.** Each protein page consists of an automatically-populated summary of peptide identifications from all the experiments.

### Bioinformatics Tools

Users have ready access to a number of built-in bioinformatics tools, including BLAST, ClustalW (Fig. 11), and InterProScan (Fig. 12), within SPW. They can use these tools to find similar proteins, identify important regions of a sequence, estimate the functions of unknown entities, or perform other analyses on their data. Analysis is run remotely via Web Services provided by the European Bioinformatics Institute (EBI) [6], enhancing the scalability of the wiki to support many simultaneous jobs.

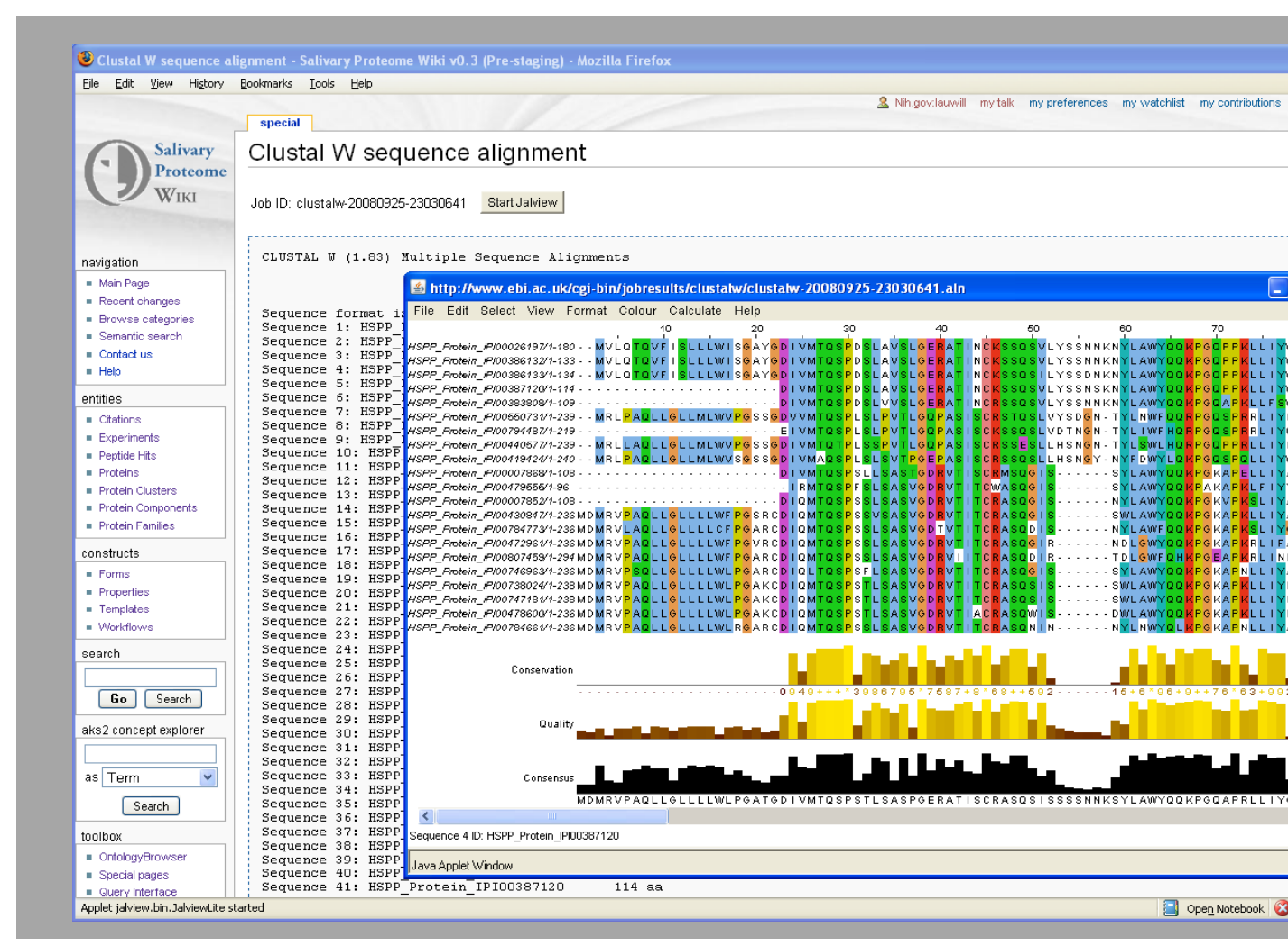


**FIGURE 11.** Multiple sequence alignment of proteins with the same set of peptide hit evidence, shown in the Jalview editor.

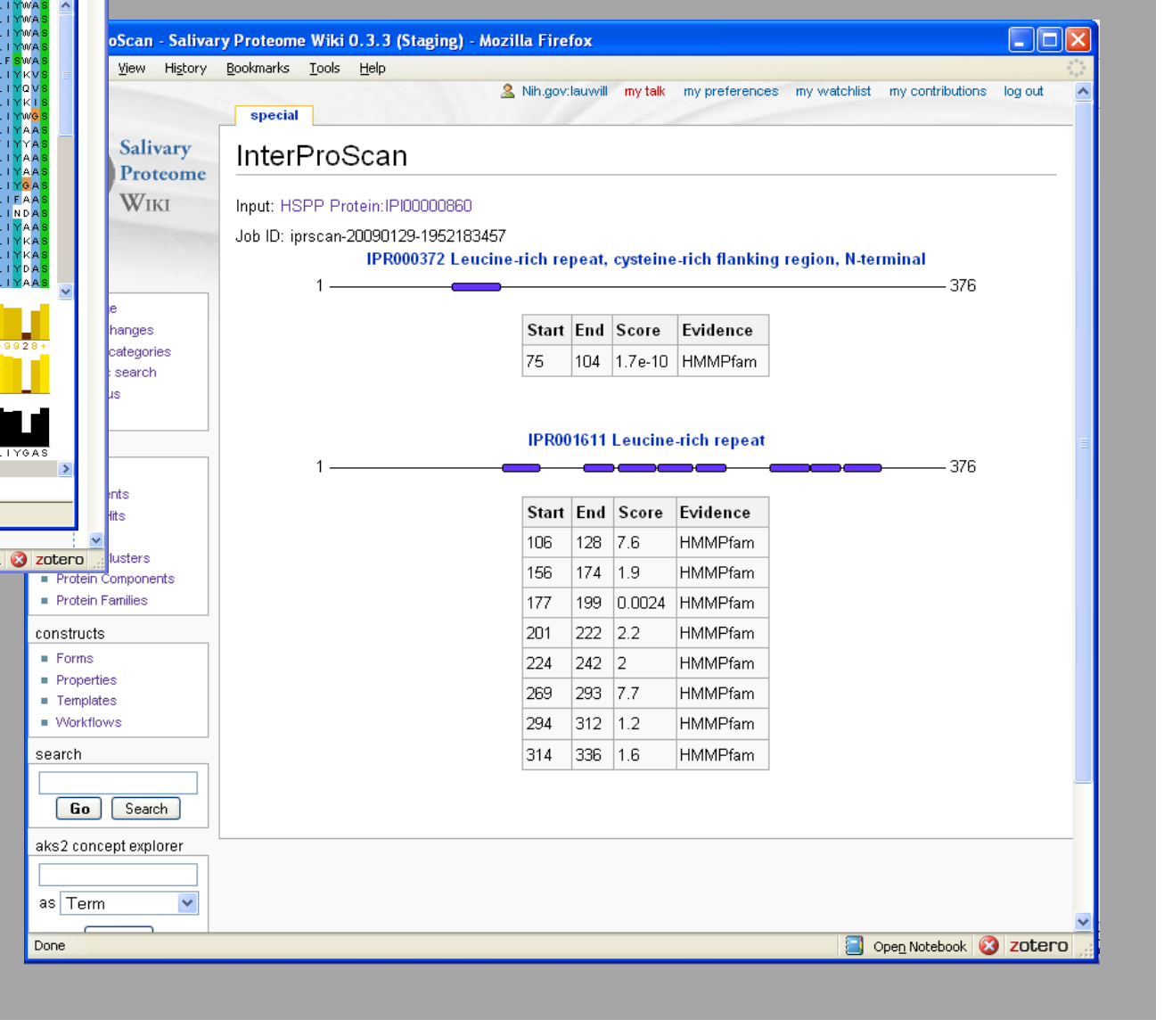
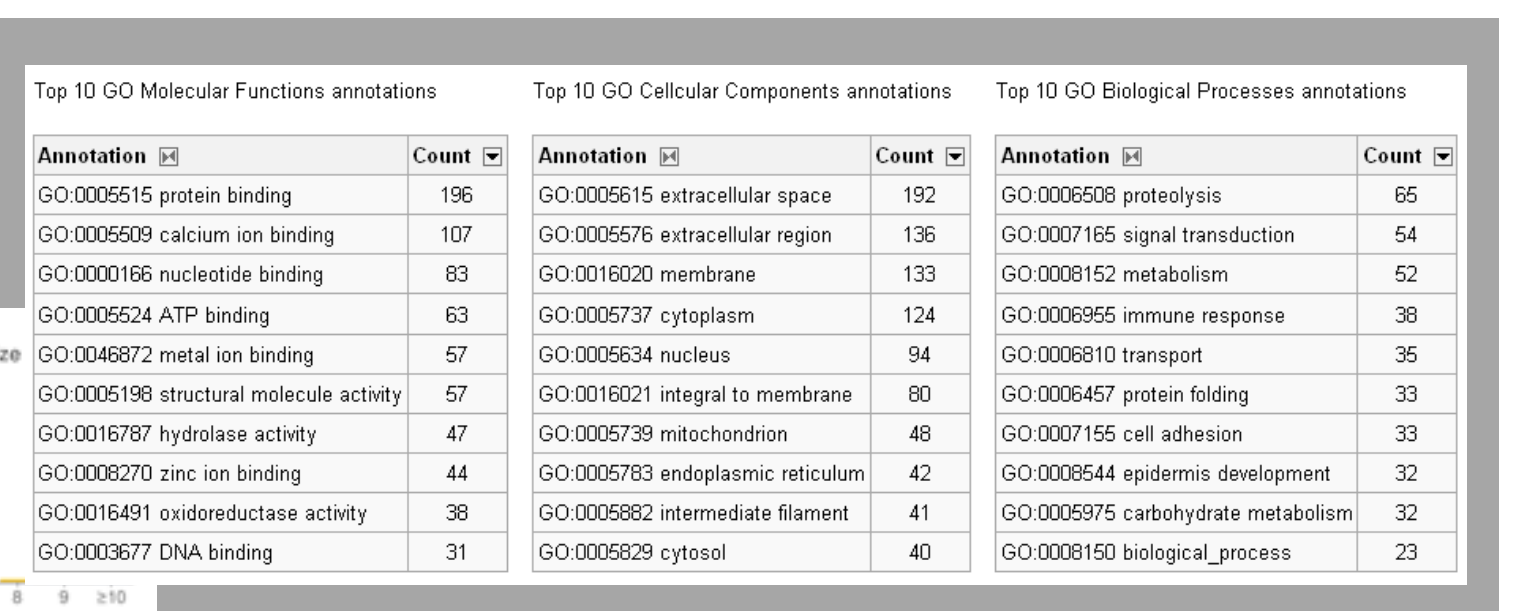


**FIGURE 12.** Results of an InterProScan analysis, highlighting significant regions in the input protein sequence.

## Results

An effective knowledge management solution is needed to handle an ever increasing volume of data generated from high-throughput experiments. The Salivary Proteome Wiki is an open resource designed to promote community-driven annotation and evolution of the human salivary proteome. By incorporating information from many relevant data sources, and utilizing Web 2.0 as well as Semantic Web technologies, the collective intelligence of the community can be exploited to the fullest extent in distilling the data into useful knowledge.

**FIGURE 13.** Top 10 most frequently used GO annotations in the salivary proteome.



Top 10 GO Molecular Functions annotations

| Annotation | Count |
|---|---|
| GO:0005515 protein binding | 196 |
| GO:0005509 calcium ion binding | 107 |
| GO:0000166 nucleotide binding | 83 |
| GO:0005524 ATP binding | 63 |
| GO:0046872 metal ion binding | 57 |
| GO:0005198 structural molecule activity | 57 |
| GO:0016787 hydrolase activity | 47 |
| GO:0008270 zinc ion binding | 44 |
| GO:0016491 oxidoreductase activity | 38 |
| GO:0003677 DNA binding | 31 |

Top 10 GO Cellcular Components annotations

| Annotation | Count |
|---|---|
| GO:0005615 extracellular space | 192 |
| GO:0005576 extracellular region | 136 |
| GO:0016020 membrane | 133 |
| GO:0005737 cytoplasm | 124 |
| GO:0005634 nucleus | 94 |
| GO:0016021 integral to membrane | 80 |
| GO:0005739 mitochondrion | 48 |
| GO:0005783 endoplasmic reticulum | 42 |
| GO:0005882 intermediate filament | 41 |
| GO:0005829 cytosol | 40 |

Top 10 GO Biological Processes annotations

| Annotation | Count |
|---|---|
| GO:0006508 proteolysis | 65 |
| GO:0007165 signal transduction | 54 |
| GO:0008152 metabolism | 52 |
| GO:0006955 immune response | 38 |
| GO:0006810 transport | 35 |
| GO:0006457 protein folding | 33 |
| GO:0007155 cell adhesion | 33 |
| GO:0008544 epidermis development | 32 |
| GO:0005975 carbohydrate metabolism | 32 |
| GO:0008150 biological_process | 23 |

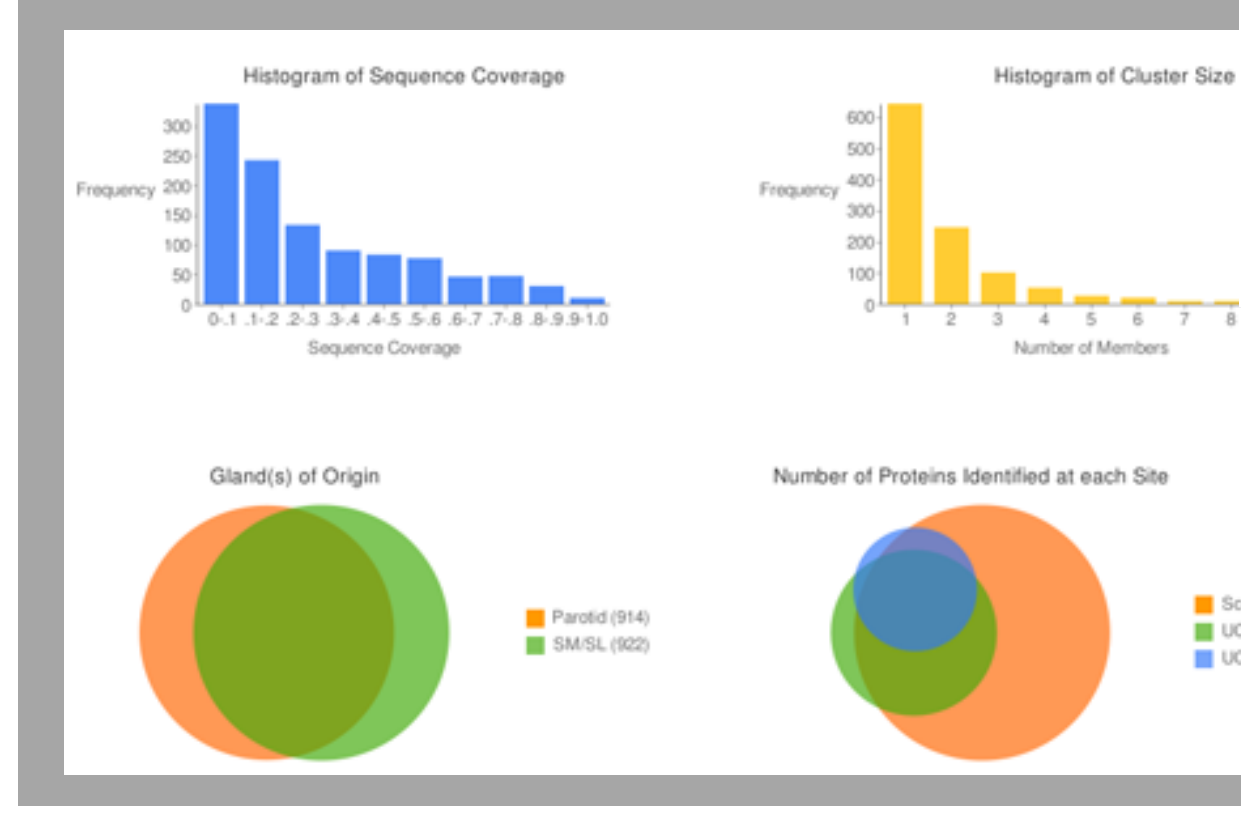


**FIGURE 14.** Summary charts generated from semantic annotations give users an overview of the information stored in the database.

A majority of proteome projects, such as the Human Salivary Proteome Project, involve researchers from across geographical locations. Many hurdles, such as issues with data integration, have to be overcome to maximize the productivity of a large-scale collaboration. In addition to being a central repository for the experimental data produced by the research groups, SPW provides a collaborative model for exchanging ideas and discovering new knowledge within a single platform (Figs. 13-14).

## References

1. Denny P, *et al.(2008)* The proteomes of human parotid and submandibular/sublingual gland salivas collected as the ductal secretions. *J. Proteome Res.7:1994-2006.*
2. Jones P, and Cote R. (2008) The PRIDE Proteomics Identification Database: Data Submission, Query, and Dataset Comparison. *Methods Mol Biol. 484:287-303.*
3. Kandasamy K, *et al. (2008)* Human Proteinpedia: a unified discovery resource for proteomics research. *Nucleic Acids Res. 37:D773-81.*
4. BioAlma. Alma Knowledge Server. *http://www.bioalma.com/aks2/index.php.*
5. *SPC Proteomics Tools. http://tools.proteomecenter.org/software.php.*
6. Labarga A , *et al. (2007)* Web Services at the European Bioinformatics Institute. *Nucleic Acids Res. 35:W6-11.*

**http://salivaryproteome.nidcr.nih.gov**

A service of the National Institute of Dental and Craniofacial Research| U.S. National Institutes of Health

National Institute of Dental and Craniofacial Research

CIT Center for Information Technology

Disclaimer: This research is supported in part by the National Institute of Dental and Craniofacial Research and the Intramural Research Program of the Center for Information Technology, NIH.